Summary of PMF study on the EPA simulated data

Philip K. Hopke and Xin-Hua Song

Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810

Data Description

The simulated data consisted of daily $PM_{2.5}$ samples for one year (1/1/1984 - 12/31/1984). Each sample was characterized by the concentrations of the following 50 chemical species: Al, NH_3, Sb, As, Ba, Bi, Br, Cd, Ca, CO_3^{2-}, Cs, Cl, Cr, Co, Cu, EC (Elemental Carbon), Ga, In, I, Fe, La, Pb, Mg, Mn, Hg, Mo, Nd, Ni, Nb, NO_3^-, OC (Organic Carbon), Pd, P, K, Pr, Rb, Se, Si, Ag, Na, Sr, SO_4^{2-}, S, H_2SO_4, Sn, Ti, V, Y, Zn, and Zr. The total $PM_{2.5}$ mass concentration for each sample, the analytical uncertainty and detection limit for each chemical species were also provided.

The objective was to infer the possible source profiles from the ambient concentration data using positive matrix factorization.

Positive Matrix Factorization (PMF)

Suppose **X** is a *n* by *m* data matrix consisting of the measurements of *n* chemical species in *m* samples. The objective of multivariate receptor modeling is to determine the number of aerosol sources, *p*, the chemical composition profile of each source and the amount that each of the *p* sources contributes to each sample. The factor analysis model can be written as:

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \tag{1}$$

where **G** is a *n* by *p* matrix of source chemical compositions (source profiles) and **F** is a *p* by *m* matrix of source contributions (also called factor scores) to the samples. Each sample is an observation along the time axis, so **F** describes the temporal variation of the sources. **E** represents the part of the data variance un-modeled by the *p*-factor model.

In PMF, sources are constrained to have non-negative species concentration, and no sample can have a negative source contribution. The error estimates for each observed data point were used as point-by-point weights. The essence of PMF can thus be presented as:

$$\min_{G,F} Q(X, \sigma, G, F) \tag{2}$$

where

$$Q = \left\| \frac{(X - GF)}{\sigma} \right\|_F^2 = \sum_i \sum_j \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2 \quad (3)$$

$$e_{ij} = x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj} \quad (4)$$

with $g_{ik} \geq 0$ and $f_{kj} \geq 0$ for k = 1, ...,p, and σ is the known matrix of error estimates of **X**. Thus, this is a least squares problem with the values of **G** and **F** to be determined. That is, **G** and **F** are determined so that the Frobenius norm of **E** divided by σ (point-wise) is minimized. As shown by Paatero and Tapper [1], it is impossible to perform factorization by using singular value decomposition (SVD) on such a point-by-point weighted matrix. PMF uses a unique algorithm in which both **G** and **F** matrices are varied simultaneously in each least squares step. The algorithm was described by Paatero [2].

Application of PMF requires that error estimates for the data be chosen judiciously so that the estimates reflect the quality and reliability of each of the data points. This feature provides one of the most important advantages of PMF, the ability to handle missing and below-detection-limit data by adjusting the corresponding error estimates. In the simulated data, there were some below-detection-limit values for different chemical species. As the input to the PMF program, the concentration data and the associated error estimates were constructed as follows. For the measured data (above detection limit), the concentration values were used directly, and the error estimates were built as the analytical uncertainty plus a quarter of detection limit. For the below-detection-limit data, half of the detection limit was used as the concentration value, and as the error estimate as well. This strategy [3] appeared to work well in the present study.

Results and Discussion

Principal component analysis (PCA) was first applied to the entire data (366×50, standardized) to examine possible outliers. Figure 1 shows the score plot of the 3 largest principal components that explain 78% percent of the total variance. It can be seen that there are 2 obvious outliers. Such outliers would have significant influence on the PMF solution. Thus, the robust mode was used to reduce this influence.

A critical step in PMF analysis is the determination of the number of factors. Analysis of the goodness of model fit, Q, as defined in Equation (3), can be used to help determine the optimal number

of factors. Assuming that reasonable error estimates of individual data points are available, then fitting each value should add one to the sum and the theoretical value of Q should be equal to the number of data points in the data set. However, the resulting solutions also have to make physical sense within the system being studied. In this application, there were some below-detection-limit data points, their error estimates were mainly based on the investigator judgements, thus it was reasonable for the calculated Q value to deviate from the theoretical value to some extent. Also, it is always good practice to experiment with different numbers of factors and compare the analysis results.

The imposition of non-negativity constraints on the factors decreases the rotational freedom and, in some cases, produces unique solutions with no rotational freedom. However, the rotational ambiguity is generally inherent in PMF as well. In this application, the rotational freedom existed. The acceptable rotations were determined by trial and error.

The nine-factor results after rotation are presented. Accompanying the factors, individual error estimates were also computed for all of the factor elements. With the total PM (particulate matter) mass known for each sample, multiple linear regression (MLR) was performed to regress the total mass against the factor scores. The regression coefficients were used to transform the factor profiles into those with physically meaningful unit, ng/ng, and to apportion the mass contributions among the resolved sources. Figure 2 shows the resolved source profiles, and the associated temporal variations (source contributions) of the nine possible sources are shown in Figure 3.

Factor 1 was identified as solid material combustion, indicated by the presence of Cl, EC, Pb, K, Si, Na, SO_4^{2-}, S, and Zn. The temporal variation indicated that there was a single high peak on 2/13 that was probably driven by Cl and Zn. Both of the species had high concentrations on that day.

Factor 2 was identified as area 1, typified by the presence of Al, Ca, Fe, OC, K, Si, and Ti.

Factor 3 was identified as road 2, dominated by high concentrations of EC and OC.

Factor 4 was identified as petroleum refinery, with Al, OC, Si, SO_4^{2-} and S present.

Factor 5 was identified as residual oil combustion, by the presence of Al, EC, Ni, OC, Si, S and V.

Factor 6 was identified as lime kiln, with Ca the dominant element. Other major elements are Al, Fe, K, Si, and S.

Factor 7 was identified as area 4. It has all of the chemical species of factor 2 (area 1) but OC. The temporal variation pattern is also different from that of factor 2.

Factor 8 was identified as steel sinter, characterized by high concentrations of Cl, Cr, Cu, EC, Fe, Pb, K, Si, and SO_4^{2-}. As to its temporal variation, on 1/23 there was a single high peak that was probably driven by SO_4^{2-}, Cr and Cu. The examination of the concentration time series of these 3 species indicated

that they had the variation trend similar to that of this factor.

Factor 9 was identified as asphalt roofing, mainly driven by Cs and Co, along with EC and SO_4^{2-}. There were two high peaks in the temporal variation of this factor, on 1/23 and on 2/11, respectively, which was in agreement with the concentration variation of Cs and Co.

To have a quantitative examination of the resolved factors, they were compared with the corresponding reference source profiles as shown in Figures 4 and 5. The error estimates for the resolved factors are also shown. It can be seen that 6 factors (solid material combustion, area 1, road 2, petroleum refinery, residual oil combustion and area 4) were in good agreement with the reference profiles. The other 3 factors were not, quantitatively, but they were identified by the presence of characteristic species (Ca, Al, Fe, K, Si and S for lime kiln, Cl, Cr, Cu, EC, Fe, Pb and SO_4^{2-} for steel sinter, Cs, Co, EC and SO_4^{2-} for asphalt roofing).

As one of the methods to determine the quality of the fit, the scaled residuals, e_{ij}/F_{ij} , were examined. Figure 6 and 7 shows the scaled residual frequency distributions for 16 major chemical species. It can be seen that most of them have symmetrical distributions and are within 3-unit range, indicating that the original data were modeled quite well by the nine factors.

Through MLR regression of the observed total PM mass against the factor scores, the reconstructed (predicted) PM mass for each sample was obtained as shown in Figure 8, with the average mass contribution of each factor to the total mass in Figure 9. The correlation coefficient between the predicted and observed PM mass is 0.99, indicating statistically the nine resolved factors account for the total PM mass very well. Regarding the individual source contributions (average), the sources area 1 and road 2 contributed the most to the total PM mass, the next major contributors were residual oil combustion and lime kiln. All of the other sources contributed much less.

Conclusions

The EPA simulated data were studied using positive matrix factorization (PMF). Nine factors were obtained and identified as the following possible sources: solid material combustion, area 1, road 2, petroleum refinery, residual oil combustion, lime kiln, area 4, steel sinter and asphalt roofing. Generally, the resolved factors are in good agreement with the corresponding reference source profiles. Through multiple linear regression of the total PM mass against the factor scores, the source contributions of the resolved factors were estimated. Among them, the two most contributing sources were area 1 and road 2, the next major contributors were residual oil combustion and lime kiln. The other sources contributed much less.

References

[1] P. Paatero and U. Tapper (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* 5:111-126.

[2] P. Paatero (1997) Least Squares Formulation of Robust, Non-Negative Factor Analysis, *Chemom. Intell. Lab. Syst.* 37:23-35.

[3] A.V. Polissar, P.K. Hopke, W.C. Malm, J.F. Sisler (1998) Atmospheric Aerosol over Alaska: 2. Elemental Composition and Sources, *J. Geophys. Res.* 103: 19,045-19,057.

Figure Captions

Figure 1. Score plot for the first three principal components.

Figure 2. Source profiles deduced for the simulated data set.

Figure 3. Time series of the source contributions for each of the identified sources.

Figure 4. Comparison of the PMF source profiles with the original profiles provided by EPA for

Figure 5. Comparison of the PMF source profiles with the original profiles provided by EPA for

Figure 6. Plots of the distribution of scaled residuals for Al, Ca, Cs, Cl, EC, Fe, Pb, and Ni.

Figure 7. Plots of the distributions of scaled residuals for OC, K, Si, SO_4, S, Ti, V, and Zn.

Figure 8. Plot of the predicted aerosol mass against the “measured” mass for the simulated data set.

Figure 9. Average mass contributions for each of the sources over the whole data set.

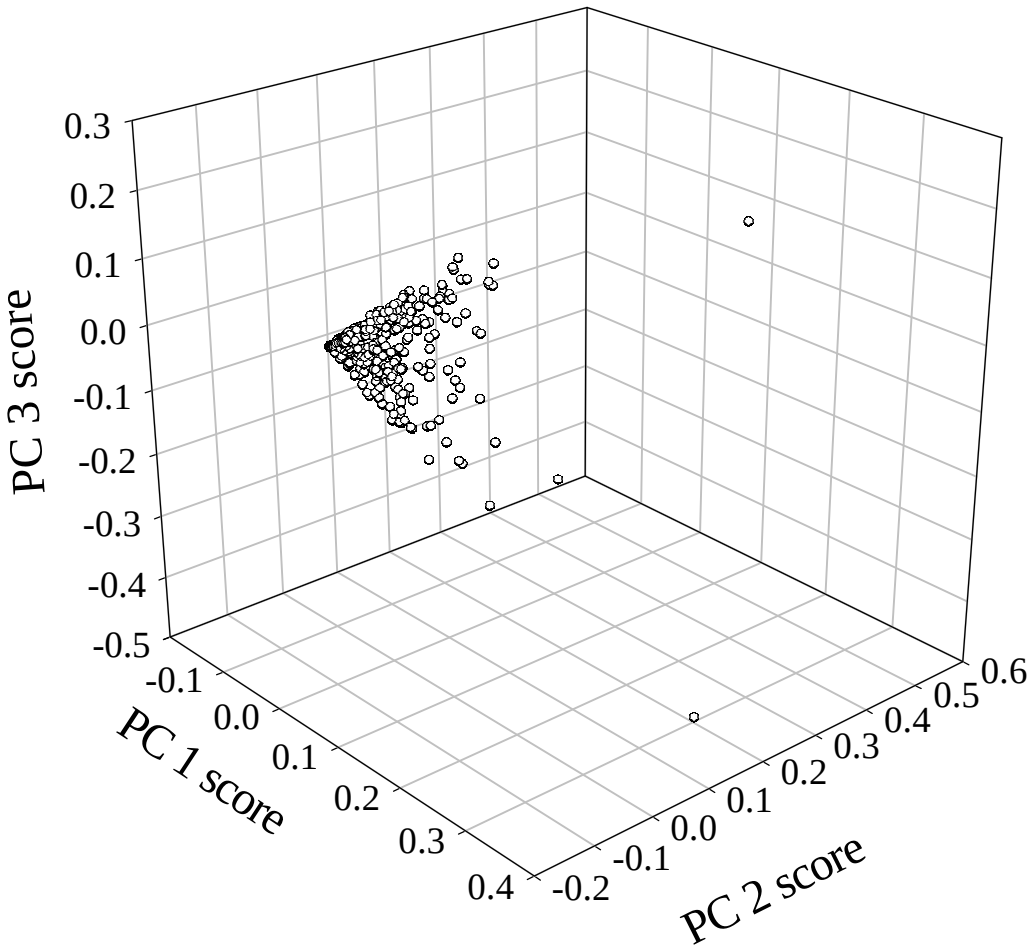
PC 3 score
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
PC 1 score
-0.1
0.0
0.1
0.2
0.3
0.4
PC 2 score
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 1.

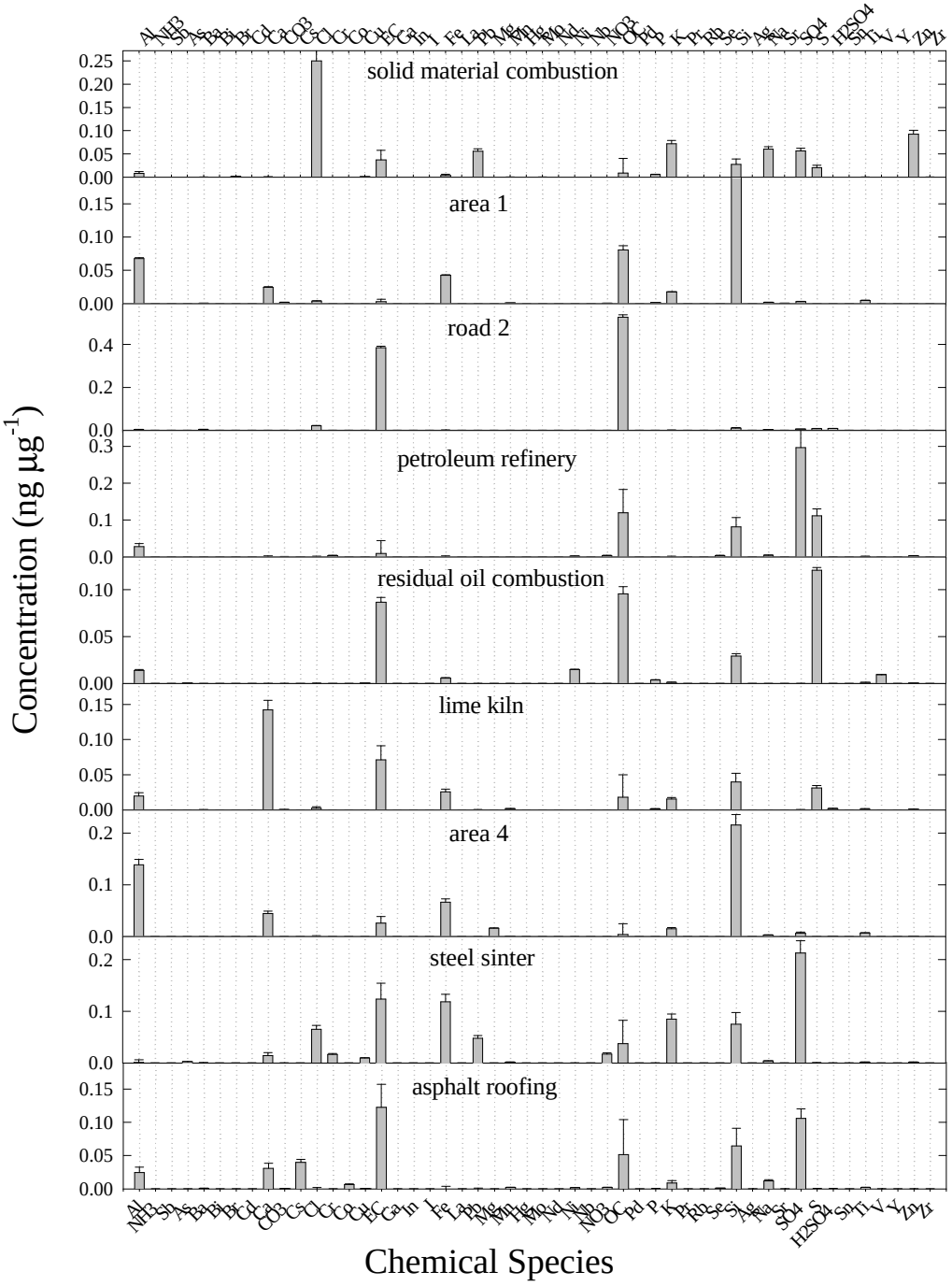
solid material combustion
area 1
road 2
petroleum refinery
residual oil combustion
lime kiln
area 4
steel sinter
asphalt roofing
Concentration (ng µg-1)
Chemical Species

Figure 2.

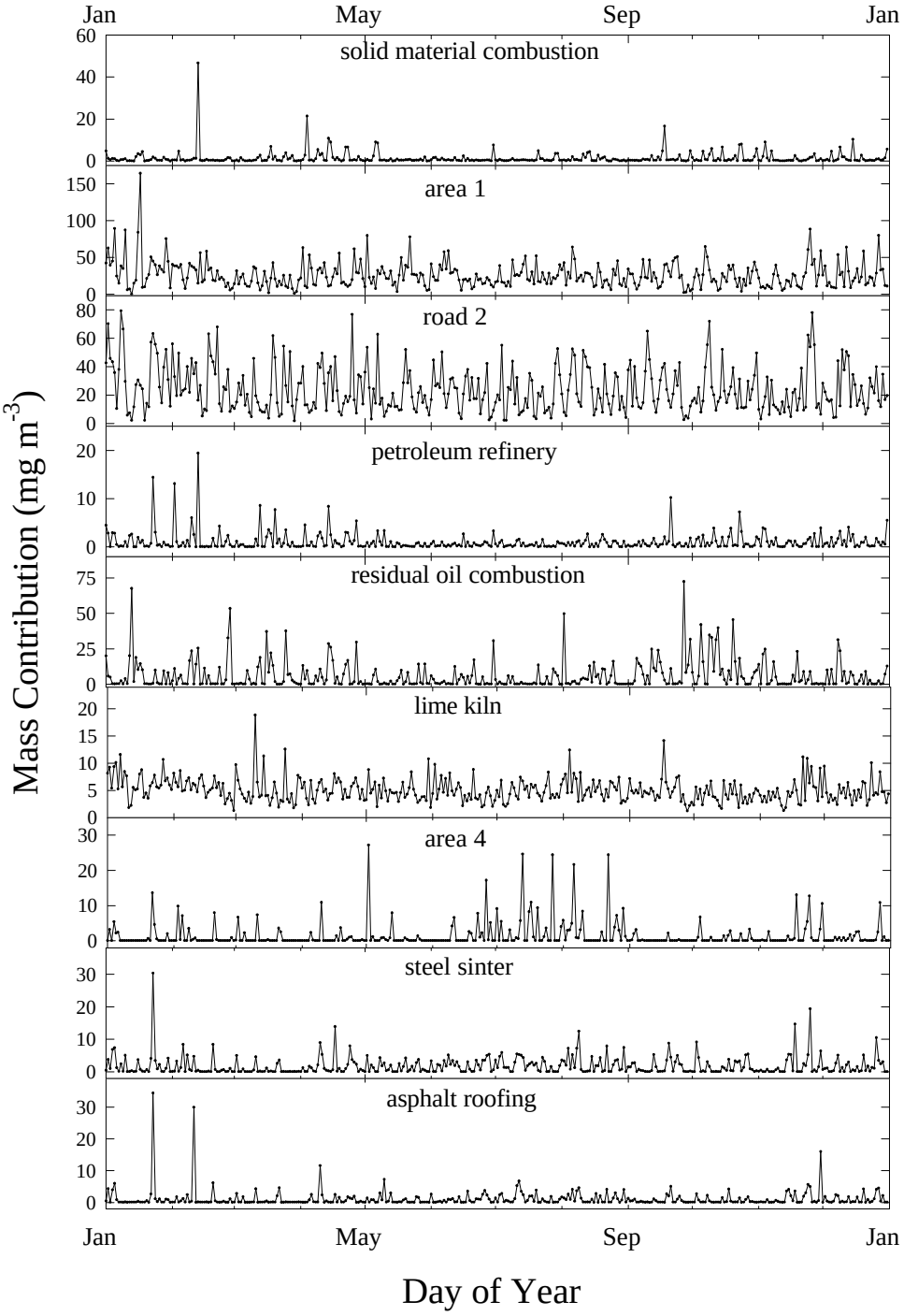
Jan
May
Sep
Jan
solid material combustion
area 1
road 2
petroleum refinery
residual oil combustion
lime kiln
area 4
steel sinter
asphalt roofing
Mass Contribution (mg m-3)
Day of Year

Figure 3.

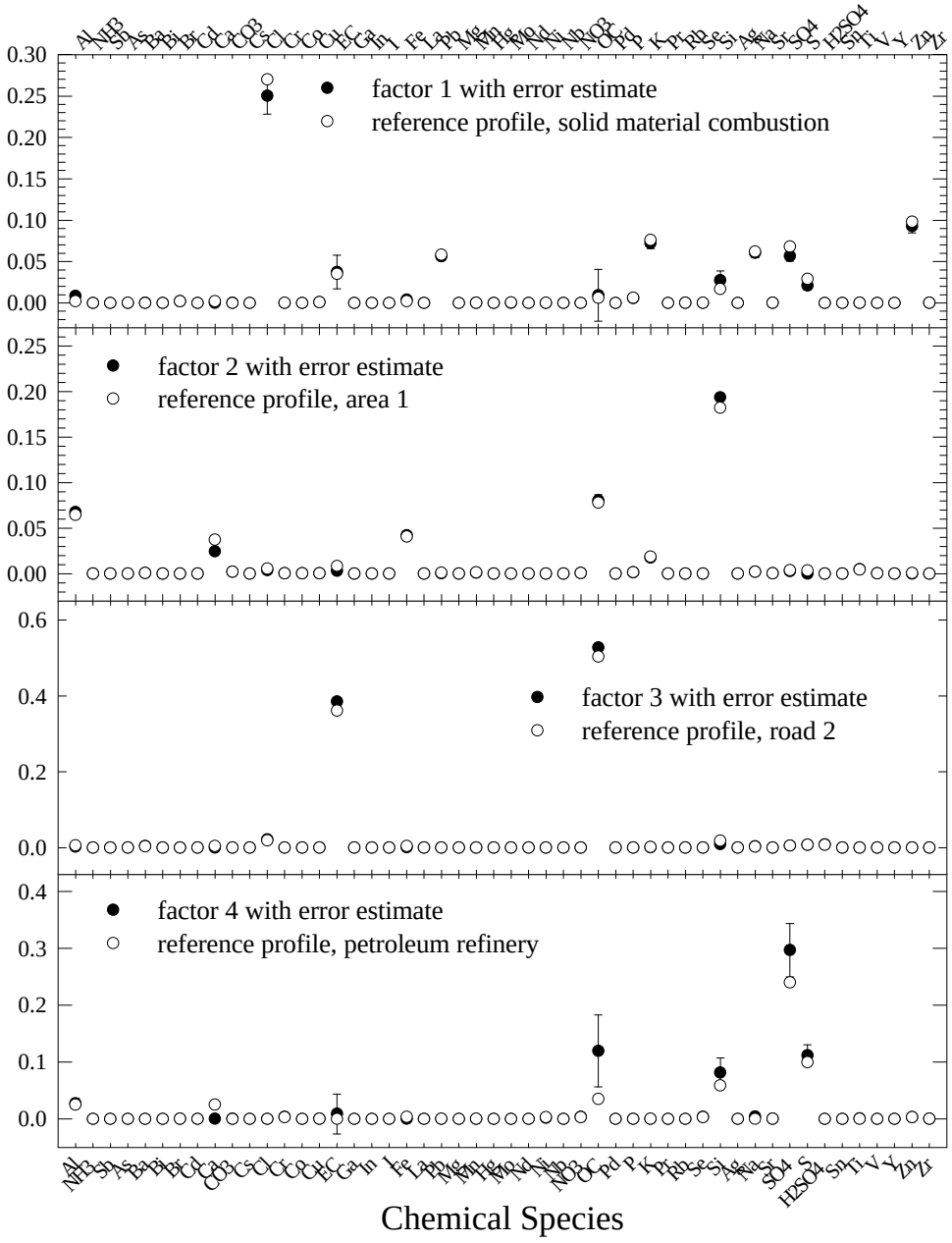
factor 1 with error estimate
reference profile, solid material combustion
factor 2 with error estimate
reference profile, area 1
factor 3 with error estimate
reference profile, road 2
factor 4 with error estimate
reference profile, petroleum refinery
Chemical Species

Figure 4.

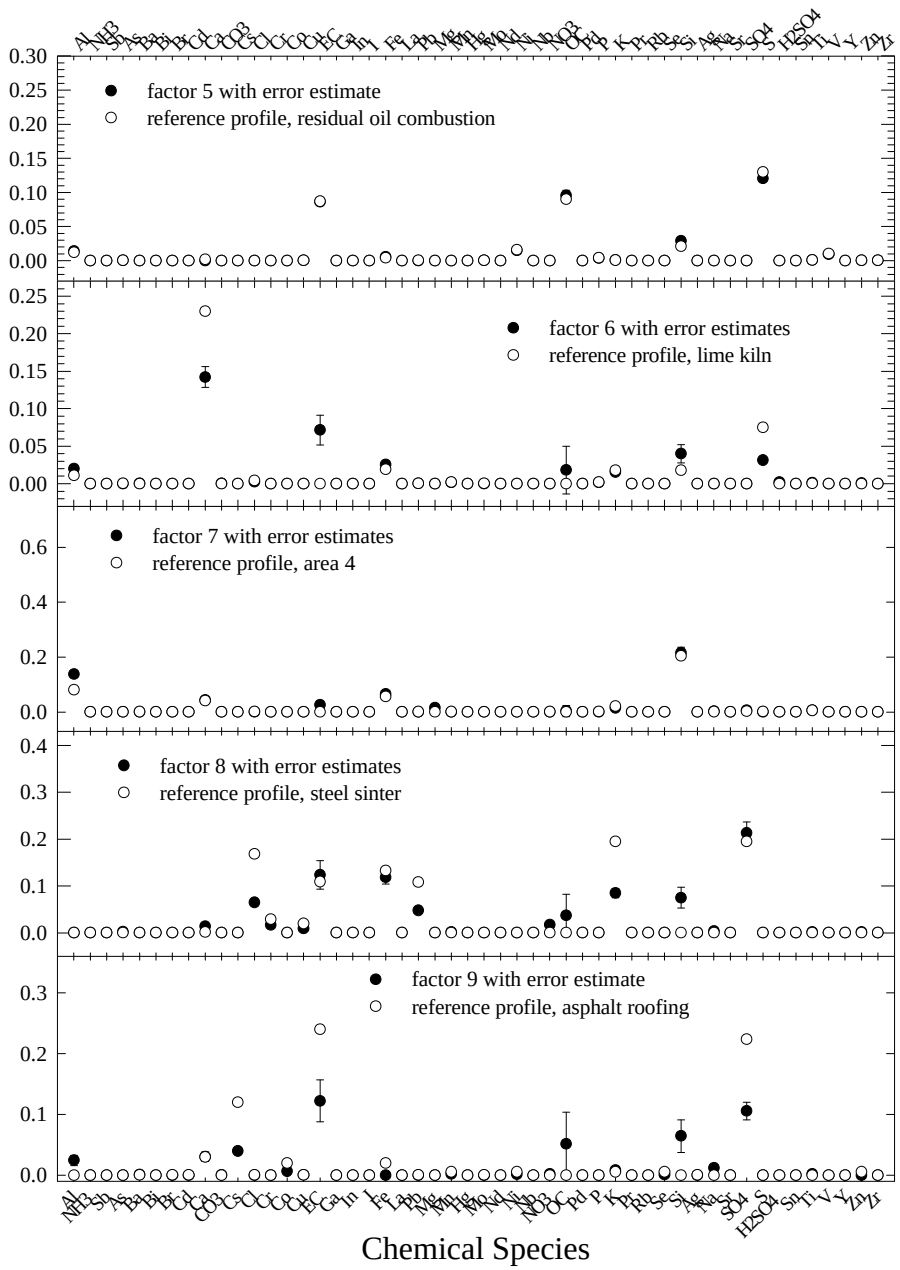
factor 5 with error estimate
reference profile, residual oil combustion
factor 6 with error estimates
reference profile, lime kiln
factor 7 with error estimates
reference profile, area 4
factor 8 with error estimates
reference profile, steel sinter
factor 9 with error estimate
reference profile, asphalt roofing
Chemical Species

Figure 5.

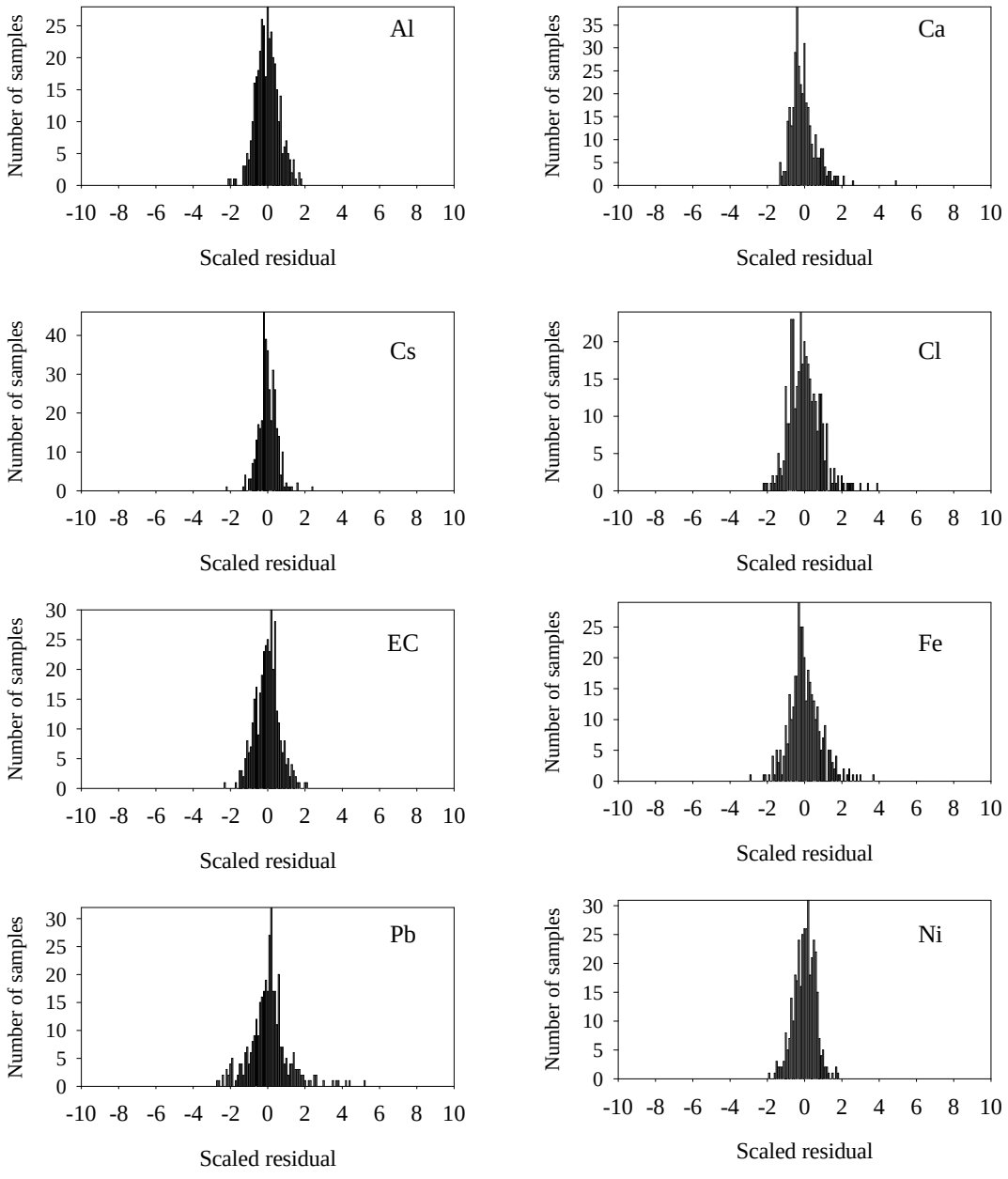

Figure 6.

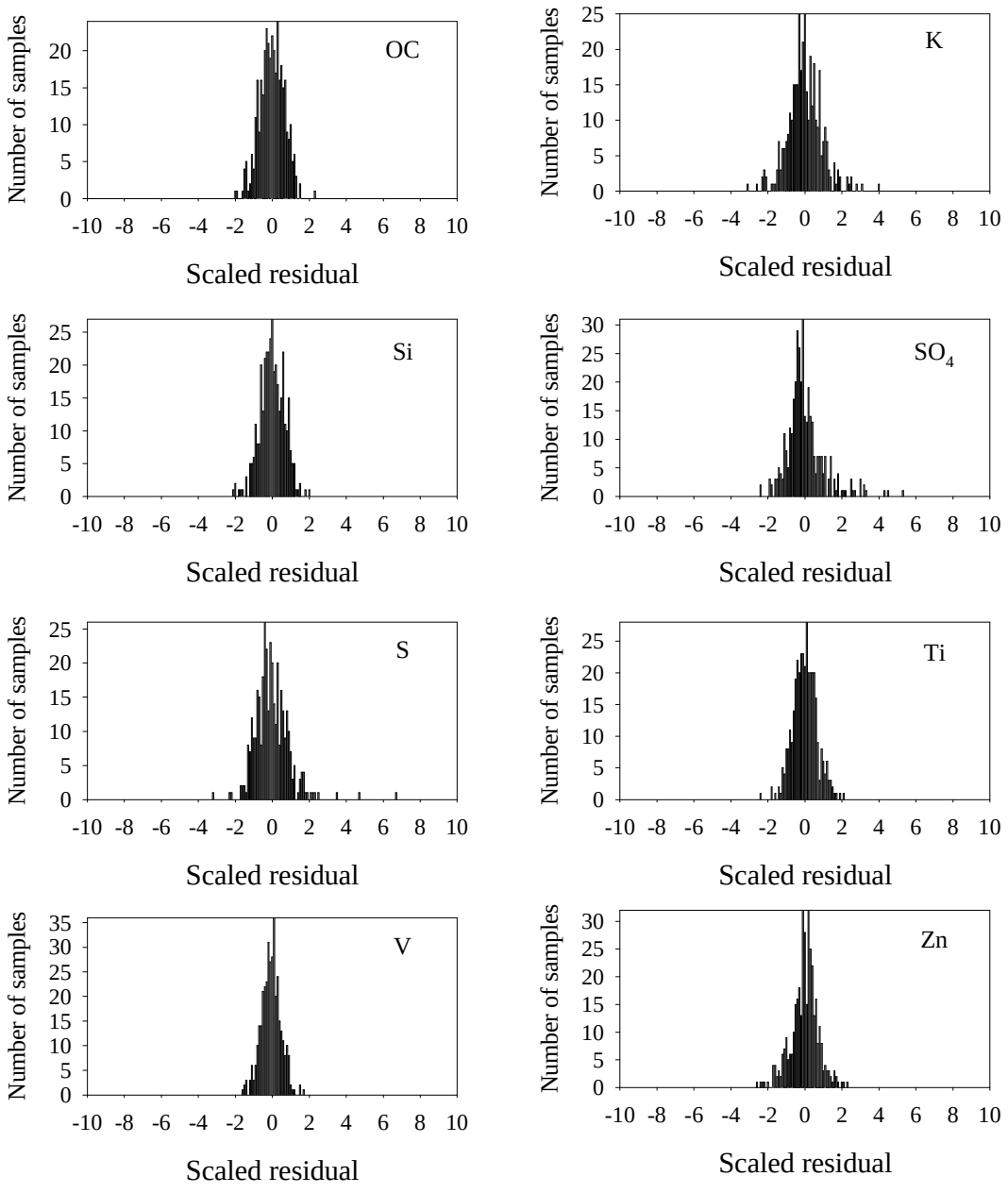

Figure 7.

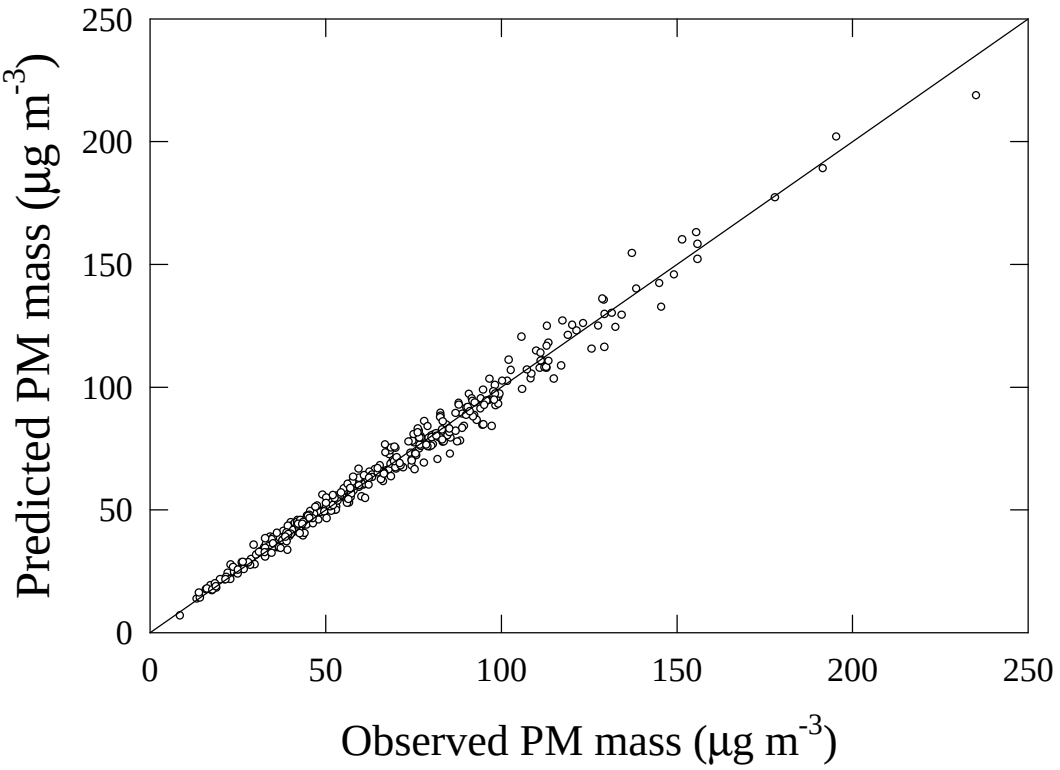
Predicted PM mass (μg m-3)
Observed PM mass (μg m-3)
0
50
100
150
200
250

Figure 8.

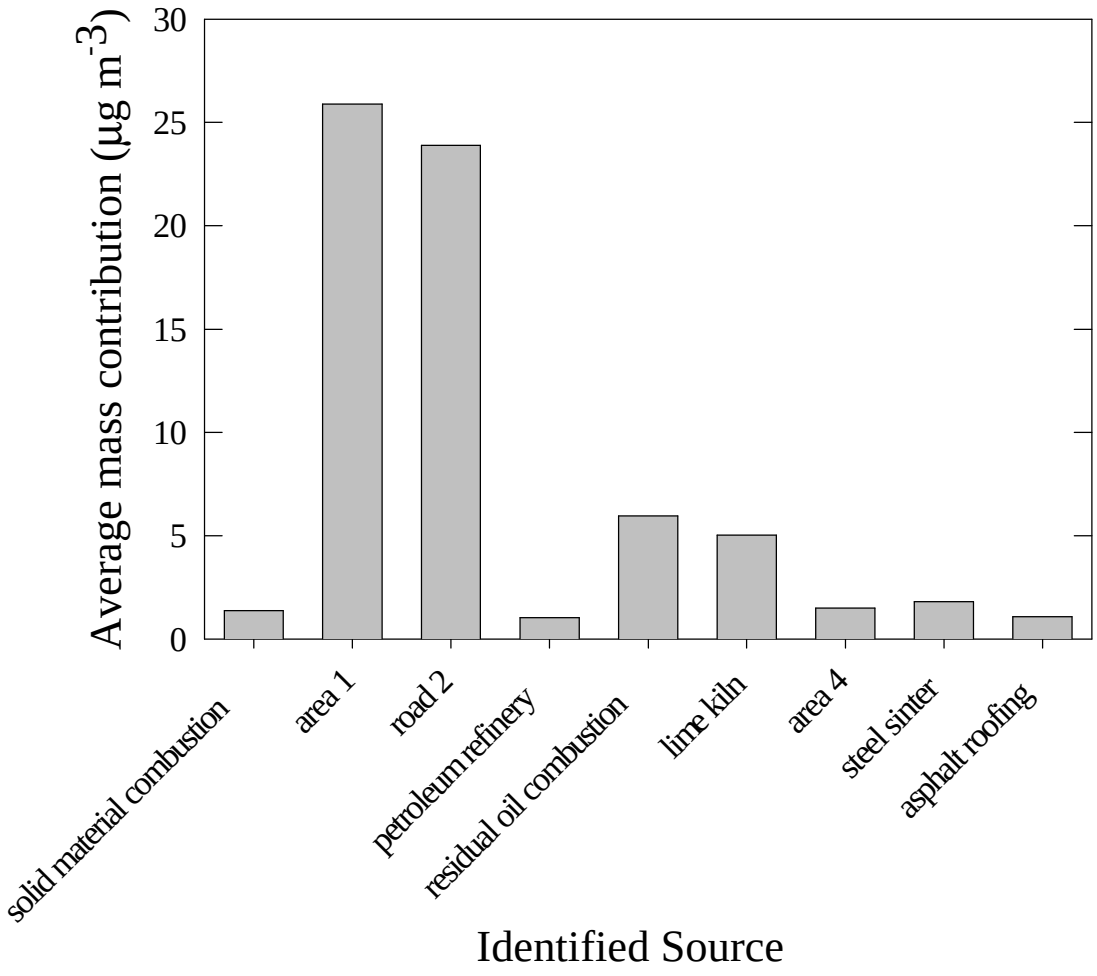
Average mass contribution (μg m-3)
30
25
20
15
10
5
0
solid material combustion
area 1
road 2
petroleum refinery
residual oil combustion
lime kiln
area 4
steel sinter
asphalt roofing
Identified Source

Figure 9.